

G.E.M.S.TM, the International Agroinformatics Alliance and G2F: from initial maintenance toward joint innovation

Kevin Silverstein, PhD

Scientific Lead, Minnesota Supercomputing Institute

Operations Manager, International Agroinformatics Alliance

5 December 2017

**College of Food Agricultural and Natural Resource Sciences
Minnesota Supercomputing Institute
University of Minnesota**



G.E.M.STM – What is it?

A novel data sharing and analysis platform to enable public-private research collaborations for innovation in agricultural production and other domain areas.



Intellectual and Institutional Assets

Our Development & Management Team (G.E.M.S)



Ph.D.
App. Econ



Ph.D.
Informatics



Software
Engineer



Ph.D. App.
Econ



Ph.D. Sci
Comp



Ph.D. Spatial
Economist



Software
Engineer



Ph.D. Spatial
Biosecurity



Ph.D. Sci
Comp



Software
Engineer



M.Sc. Data
Scientist



Ph.D. Pop
Genetics



M.Sc. Data
Scientist



Software
Engineer



M.Sc. Res.
Associate



Ph.D.
Informatics



Ph.D. Sci
Comp
Developer



Ph.D Comp
Sci
Developer



M.Sc. Res.
Analyst



Ph.D. Spatial
Data Mining

Our Collaborations



IAA 2.0 March 20-21, 2017, St. Paul MN

Our Faculty



Plant
Pathologists



Economists



Breeders



Geneticists



Agronomists



Microbial
Biologists



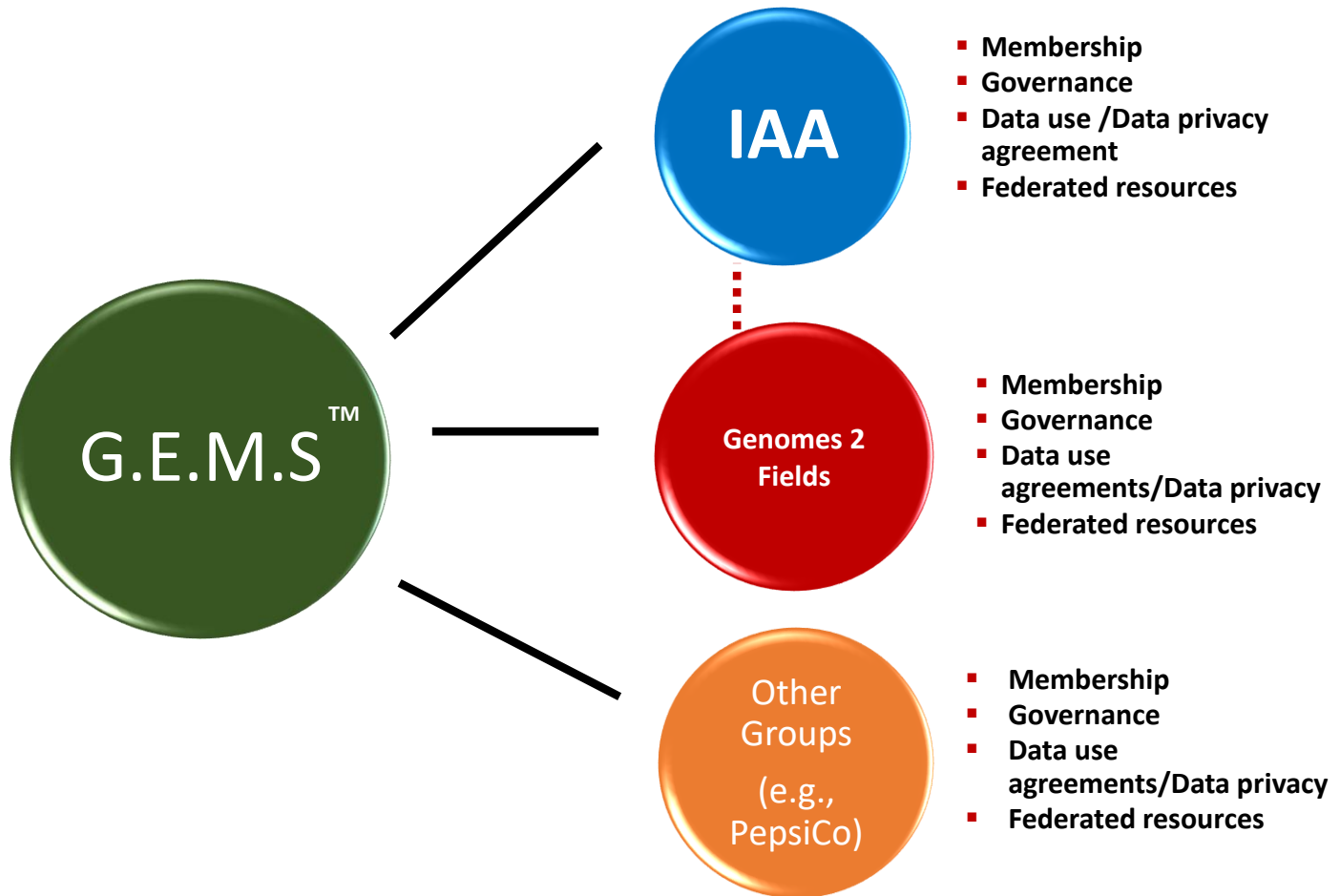
Environmental
Biologists



Engineers

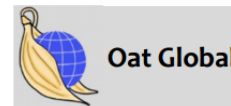


Partnerships Options



IAA (External) Partnerships

- Embrapa, Brazil
- Pepsico
- Diversity Arrays Technology (DArT/KDDART)
- CIAT (cassava, edible beans, forages, rice,)
- G2F (Genomes to Fields)
- University of Adelaide
- CSIRO, Australia
- Oat Global
- Stellenbosch University, South Africa



-
- GRDC (Grains Research Development Corporation), Australia
 - CIMMYT (corn, wheat, socio-economics, genetic resources, IT)
 - MN Department of Agriculture
 - University of Western Australia
 - CGIAR (Big Data Initiative)
 - CREEF, Phenotyping Center, Canada
 - CIP (potatoes, sweet potatoes)



G.E.M.STM – Core Features

GEMShareTM

A research-enabling, federated data storage and sharing platform

- **Security:** Appropriate levels of security (data encryption at rest; authentication with home institution's credentials; and secure infrastructure)
- **Access Control:** Data owners control access to their data, in recognition of the proprietary nature of much of the data
- **Access Levels:** Different levels of access [single organization; set of organizations; and publicly shared (open) data and analytical tools]
- **Discovery:** Discoverability of data through metadata alone
- **Transfer:** Secure data transfer over both high speed networks between reliable endpoints and high latency networks to less reliable endpoints

G.E.M.STM – Specialized Features


GEMToolsTM

An ever-expanding data documentation, cleaning, harmonizing and analysis toolkit

- provide access to best in class hardware and software libraries
- accommodate different programming languages
- offer a range of analysis styles (novice to sophisticated)

GEMSTools – Analysis Interface (Expert)

 R_iaa_demo Last Checkpoint: Yesterday at 11:31 AM (autosaved)

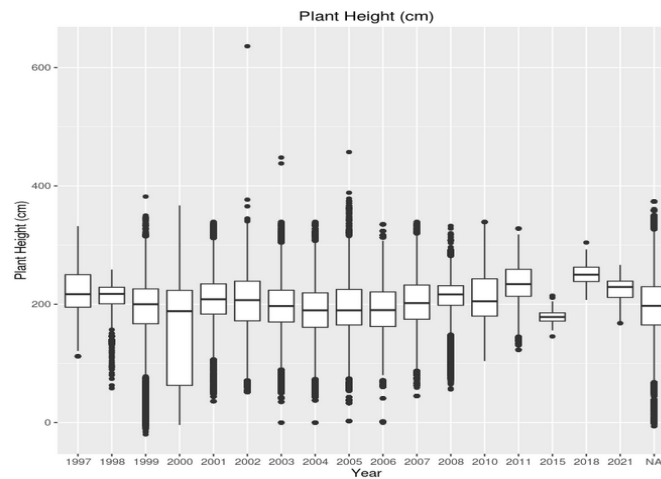
 Control Panel Logout

File Edit View Insert Cell Kernel Help

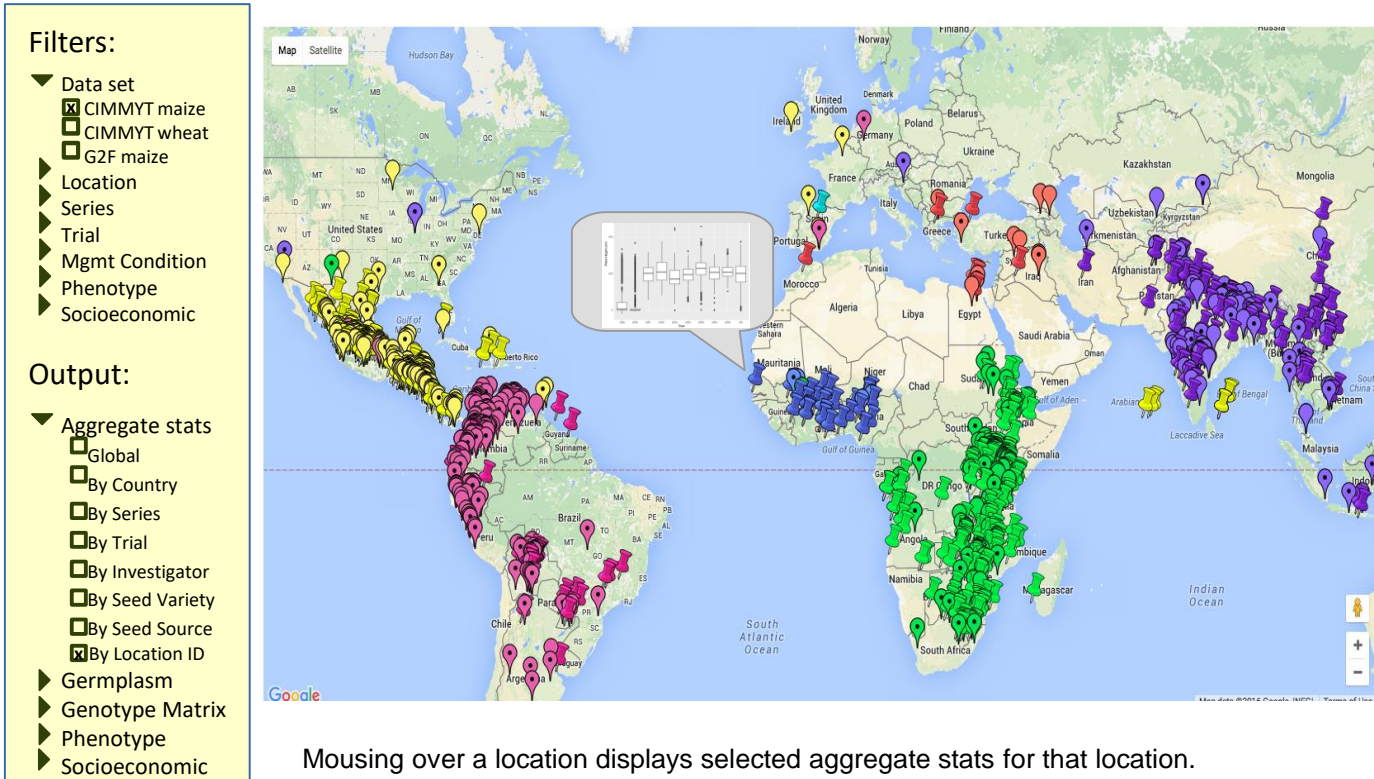
          Markdown    

Summarize plant heights globally

```
In [13]: # Query 3: aggregate data on these three features globally
p <- ggplot(data=sub_tbl, aes(factor(year), value)) + geom_boxplot()
p+ggtitle("Plant Height (cm)") + labs(x="Year", y="Plant Height (cm)")
```



GEMSTools – Analysis Interface (Point & Click)



Data interoperability issues and GEMTools™ solutions

Typical Data Impurities

- Nomenclature inconsistencies
- Measurement unit differences
- Erroneous and missing entries
- Outlier / physically impossible data values
- Domain-specific problems
 - Pedigree syntax
 - Genotype / Pedigree inconsistencies
 - Spatial concordance of census and mapped data
 - Spatio-temporal boundary standardization

GEMS Tools—DataCleaner

Nomenclature inconsistencies

Count	Management Conditions
1	Agricultura de conservación
1	Agricultura de Conservación
1	Agriculture Conservation
1	Conservacion Agriculture
1	Conservation agriculture
21	Conservation Agriculture
774	Low N
30	Low Nitrogen
336	Managed Low Nitrogen
20	Maize Streak Virus
4	msv
10	MSV
...	

Spatio-temporal boundary standardization

- Standardize attributes and metadata
 - Unit names
 - Unit codes
 - Unit levels
- Clean boundaries
 - Remove overlaps
 - Check geospatial consistency
- Build spatial and temporally explicit boundary geodatabase
 - Document boundary changes with acceptable spatial and temporal accuracy
- Store with standardized GEMS code and metadata with unique identifier for each unit

Spatial Concordance of Census and Mapped Data

Year: 1920
Municipalities: 1,303



Measurement Unit Differences

- **Grain Yield**

- CIMMYT

- Adjusted to 12.5% grain moisture
 - Units of kg/hectare

- G2F

- Adjusted to 15.5% grain moisture
 - Units of bushels/acre

- **Plant Height**

- CIMMYT

- Measured to insertion of first tassel branch

- G2F

- Measured to flag leaf (or to top of plant for TX)

Dynamic Metadata Mashup Model—DM³



- **EML (Ecological Metadata Language):** experiment, investigator, institution, organism specimens and taxonomy.



- **OBI:** sequencing, library preparation, and sequence processing

- **ENVO & XEO/XEML:** environmental features and habitats

- Planteome.org **(TO & CO):** plant phenotypic traits (TO) across many individual crop ontologies (CO)



- **PATO:** plant phenotypic qualities

- **AGRO** agronomic practices and techniques



- **OGC standard ISO19115-2, FGDC and Dublin Core:** geospatial

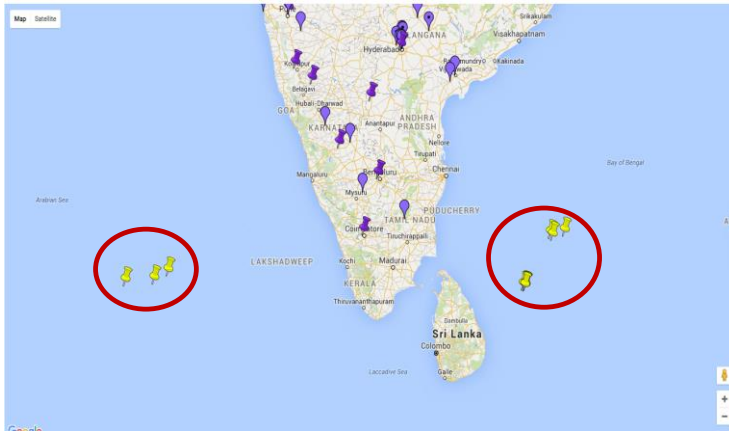
- **Broad Vocabularies**

- **AGROVOC (FAO):** including food, nutrition, agriculture, fisheries, forestry, environment etc. Translated in 27 languages.

- **ICASA (AgMIP)**

GEMSTools — DataCleaner

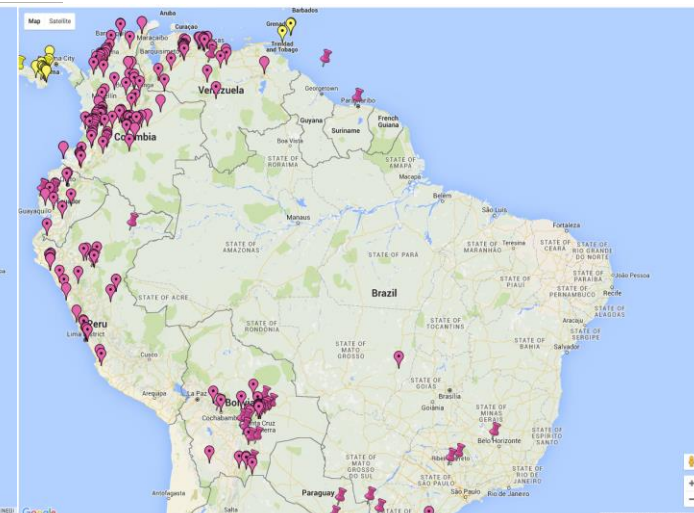
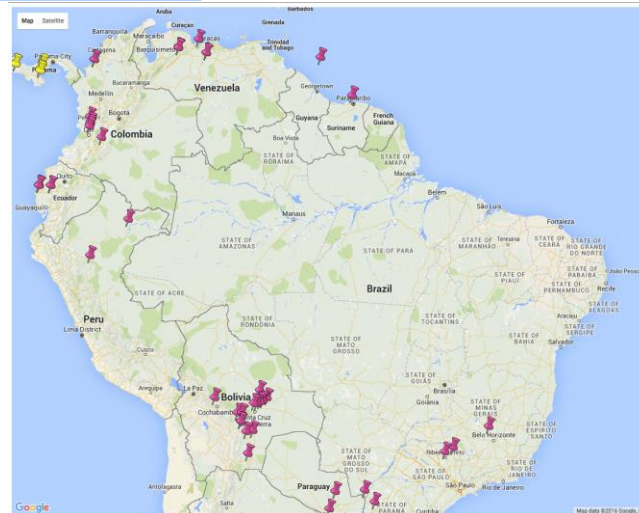
Erroneous and missing entries



Before

After

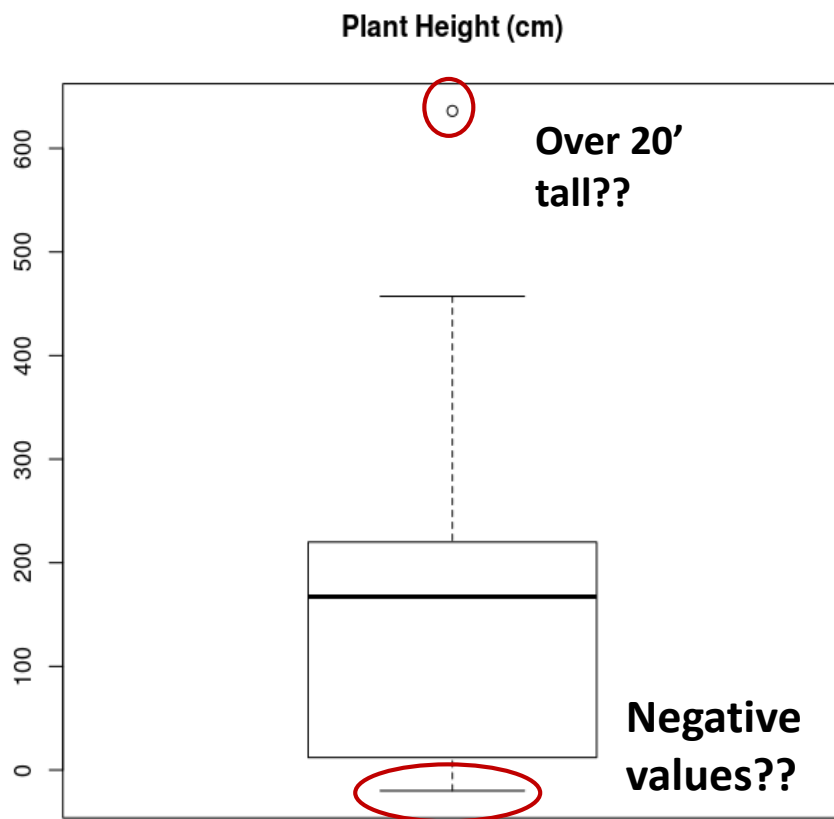
Correcting errors



Imputing missing lat/long values

GEMS Tools—DataCleaner

Outlier / physically impossible data values



Jack in the corn stalk: Record 45-foot-tall corn plant created by New York breeder



GEMSTools – PedTools

Pedigree Syntax Cleaner

Pembina*6 /2/ Thatcher*3 / Transfer

PMB*6//THA*3/TRANSFER

Pembina*6 /3/ Thatcher*2 /2/ Marquis*6 / Red Egyptian

PMB*6/3/THA*2//MRQ*6/REGY

RL4205

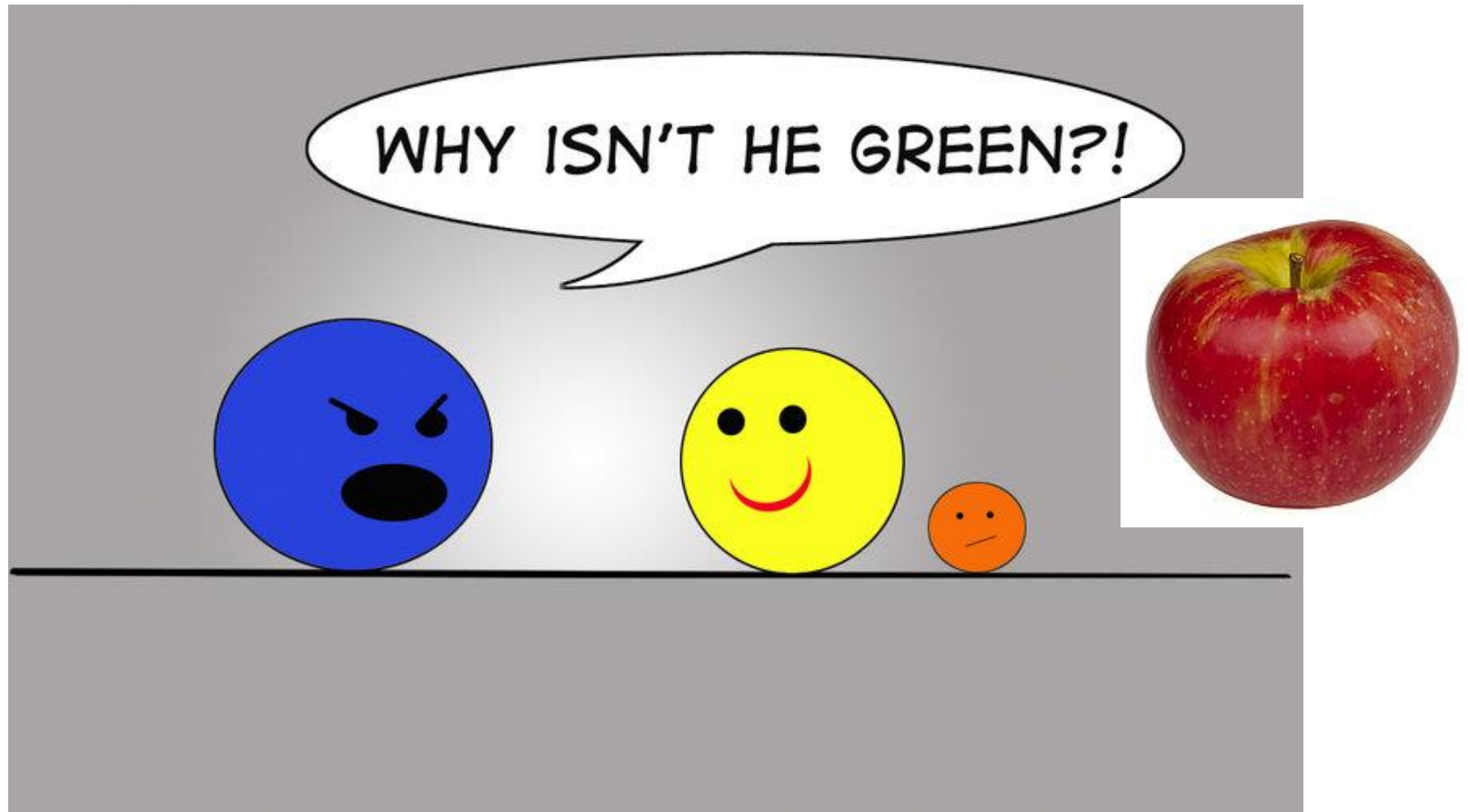


Nested pedigree parsing can be tricky:

Alex: Waldron /5/ (**RL4205, Pembina*6 /2/ Thatcher*3 / Transfer /4/ Pembina*6 /3/ Thatcher*2 /2/ Marquis*6 / Red Egyptian**) /9/ (ND496, Waldron /8/ (ND269, Conley /7/ (ND122, Maria Escobar / Newthatch /6/ Kenya 338AA /5/ Lee /4/ (N1831, Mida /3/ (N1530, (H-44 / Ceres, N1349-15) /2/ Thatcher))))))

Genotype / Pedigree Inconsistencies

Genotype vs Pedigree with Jim Luby & Nick Howard



Who's Honeycrisp's REAL father?? Inquiring minds want to know.
And now we do!!

GEMSTools™— Machine Aided Data Cleaning

- Modular code to address each cleaning issue
 - Work on specific problem (e.g., maize field trial data)
 - Write code to automate much of the cleaning
 - Apply to new crops or new datasets
 - G2F vs CIMMYT vs PepsiCo (nomenclature cleaning)
 - Maize, wheat, soybean, apples (pedigree cleaning)
- Rule-based techniques, Natural Language Processing, and some Deep Learning methods
- Converge toward real-time feedback on cleaning

IAA and G2F: current agreement

- **Iowa Corn has provided funds for**
 - \$4,000 membership to IAA
 - 10% effort Kevin Silverstein
 - 100% effort Christina Poudyal, new data manager
- **Initial plans**
 - Work with Naser to collect, clean, summarize and distribute 2017 trial data prior to March community meeting
 - Use existing infrastructure for this year's data

IAA and G2F: medium-term ideas

- **Talk with collaborators to improve existing operations**
 - Streamline data collection and submission?
 - Enable real-time data cleaning upon upload?
 - Enable real-time analysis and year-to-year comparison?
- **Enable analysis of transformative technologies**
 - Real-time incorporation of weather/environment data?
 - Analysis of UAV and remote-sensing data for each field?
 - SOWs for in-season predictive analytics?

Thanks

G.E.M.S / IAA URL – Under Construction!



Dr. Norman E. Borlaug
1914 - 2009
University of Minnesota
B.S. Forestry 1937
M.S. Plant Pathology 1941
Ph.D. Plant Pathology 1942
*"If you desire peace, cultivate justice, but at the same time cultivate the
justice to prevent more blood wherever there will be no peace."*
Nobel Peace Prize (1970)
Presidential Medal of Freedom (1971)
National Academy of Sciences, the National Medal of Science (2006)
Congressional Gold Medal (2007)

Basic Development Principles

- Leverage actively developed open source software and libraries
- Contribute back to open source development
- Build new communities of developers and users when none exist
- Prepare to throw stuff away

Open Source Tools Supporting G.E.M.S

Postgres - MIT
Jupyter - BSD 3.0
Django - BSD 3.0
pyCSW - MIT
Globus - Apache 2.0
Apache Spark - Apache 2.0
Geotrellis - Apache 2.0

Docker - Apache 2.0
PostGIS extensions - GPL 2.0
Puppet - Apache 2.0
Conda - BSD
R - GPL
Scala - BSD
CentOS

G.E.M.STM – Business Models

Three different business models to accommodate the main types of collaborations, each with its own financial plan for long-term sustainability.

Software as a Service (SaaS): On-line access to the G.E.M.STM platform, data sets, and apps managed by MSI or a federated partner.

Open Core: Access to the G.E.M.STM platform software, which may be run on user defined infrastructure from a laptop to a cloud source provider.

Data as a Service (DaaS): Users have access to the sharable G.E.M.STM data but otherwise use their own infrastructure and applications for analysis.